www.bsc.es

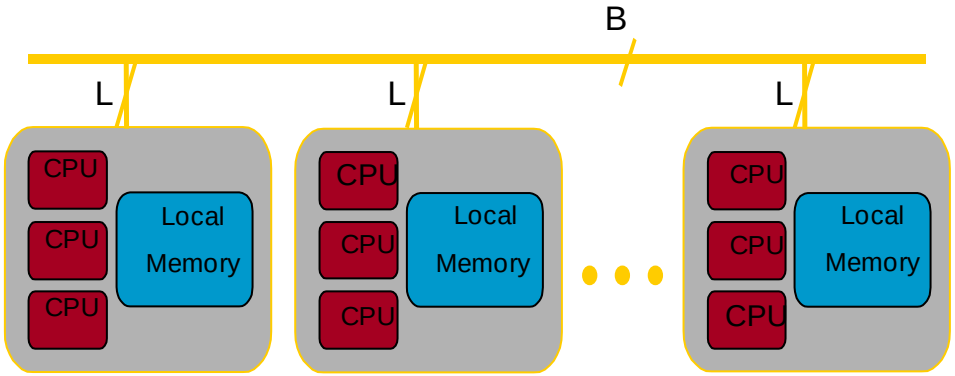
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

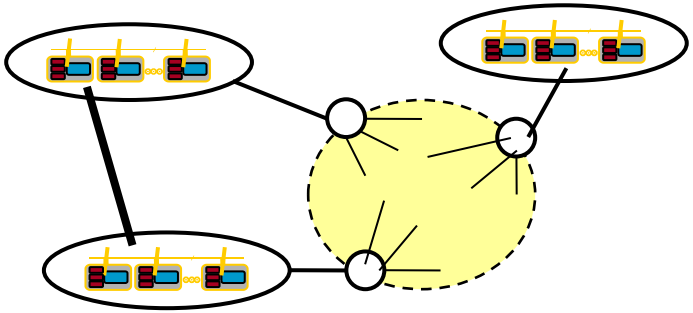
Performance analysis with BSC-Tools DIMEMAS simulator

Juan Gonzalez – juan.gonzalez@bsc.es

DIMEMAS: Coarse grain trace driven simulator

- Key factors
  - Abstract architecture
  - Basic MPI protocols
  - No attempt to model details
- Objectives
  - Simple / General
  - Fast simulations
- Linear Components
  - Point-to-point
  - CPU / Module speed ratios
- Non-linear components
  - Synchronizations
  - Resource contention
- Network of SMPs / GRID

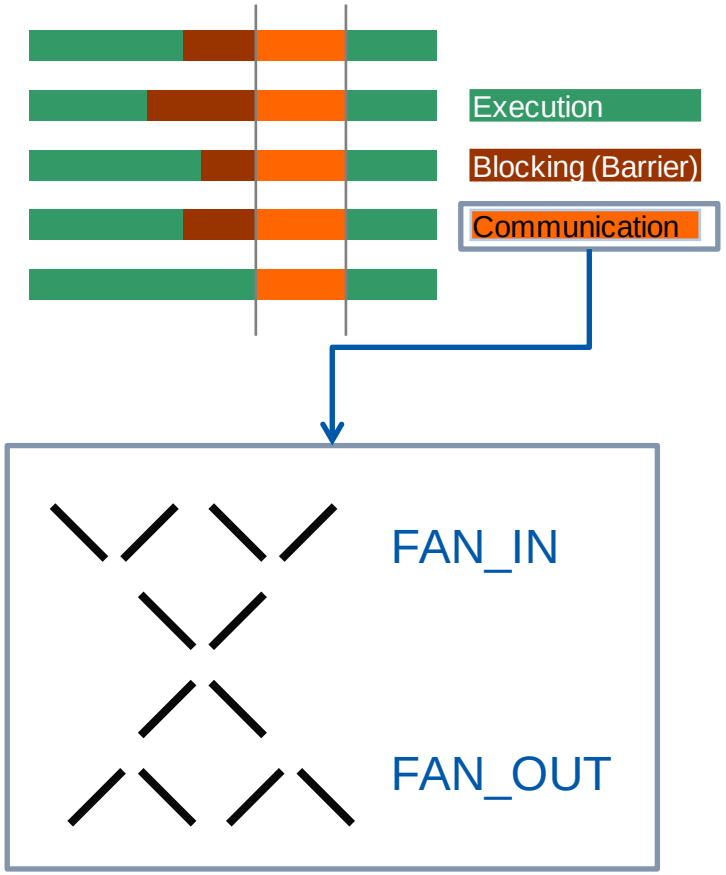




$$T = L + \frac{\text{MessageSize}}{BW}$$

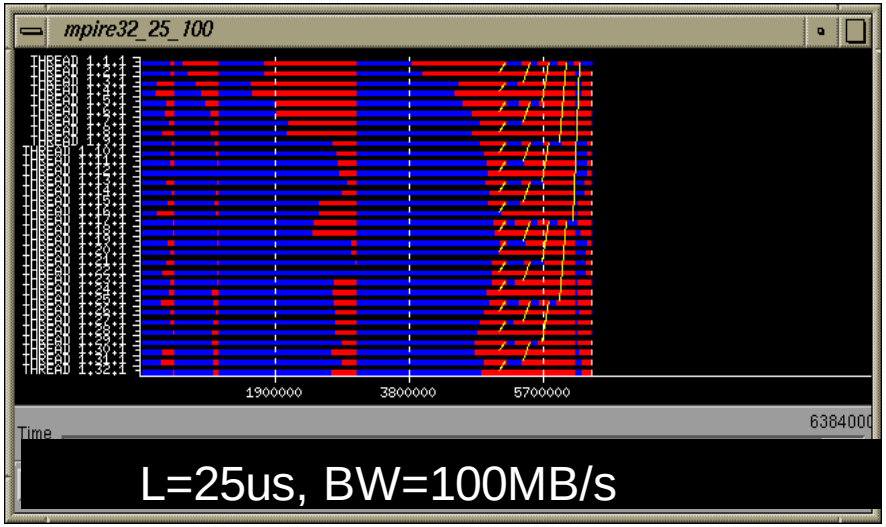
Collective communication model

- Structure
  - Barrier / FAN-IN / FAN-OUT
- Cost of communication phase
  - Generic
  - Per call
    - Model factor
      - LIN / LOG / CONST
    - Size of message
      - MIN / AVG / MAX

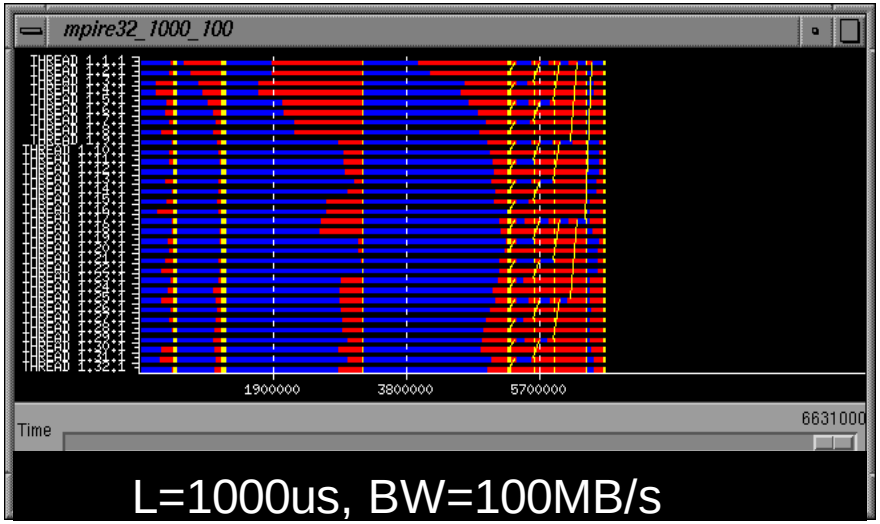


Understanding applications

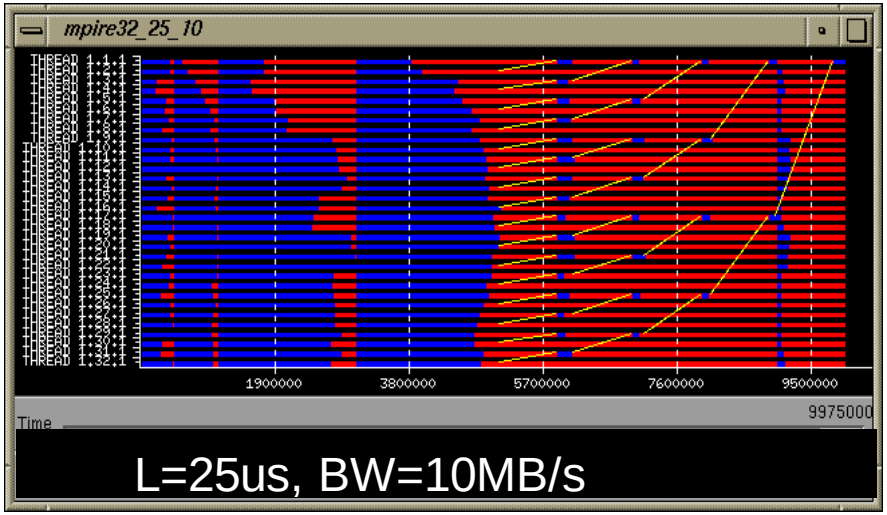
- MPIRE 32 tasks, no network contention



L=25us, BW=100MB/s



L=1000us, BW=100MB/s

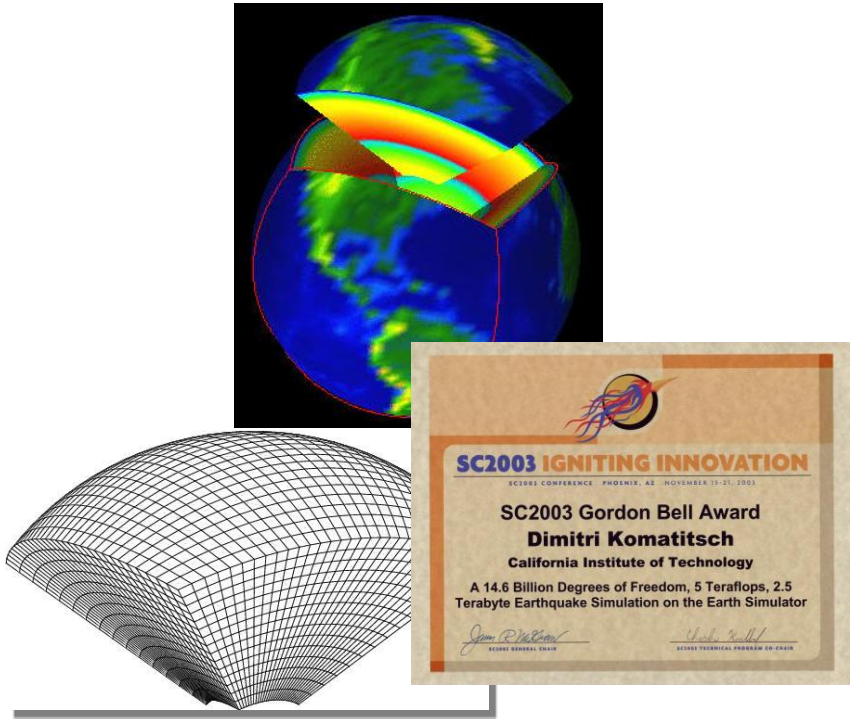


L=25us, BW=10MB/s

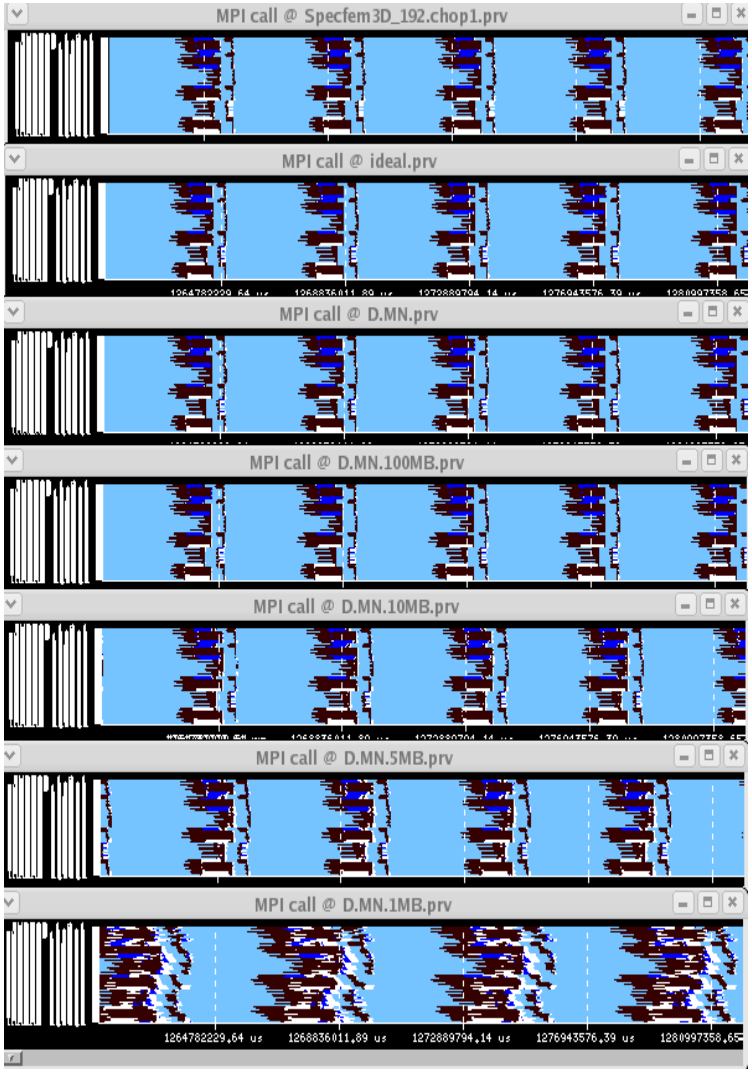
All windows same scale

Predicting performance

- SPECFEM3D
  - Do we need asynchronous communications?

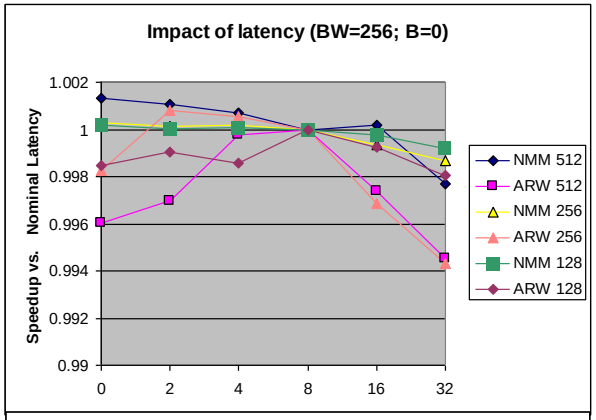


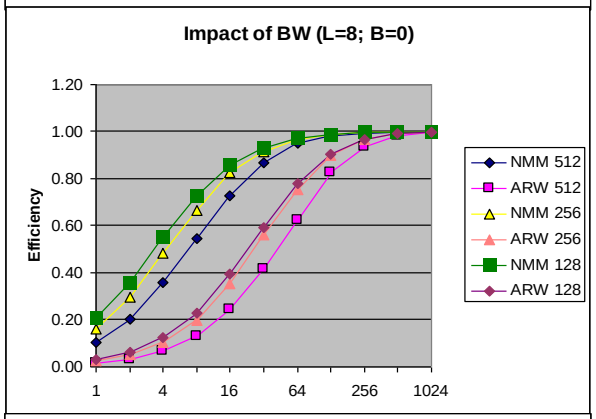
Courtesy Dimitri Komatitsch

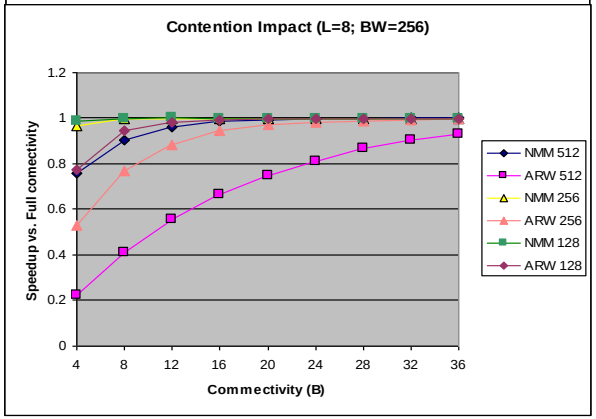


Parametric studies – Network sensitivity

- WRF
  - Iberia 4Km, 4procs/node
  - NMM / ARW models
- None sensitive to latency
- NMM
  - Bandwidth (BW). 256MB/s
  - 512 proc. → sensitive to contention
- ARW
  - BW. 512MB/s
  - Sensitive to contention

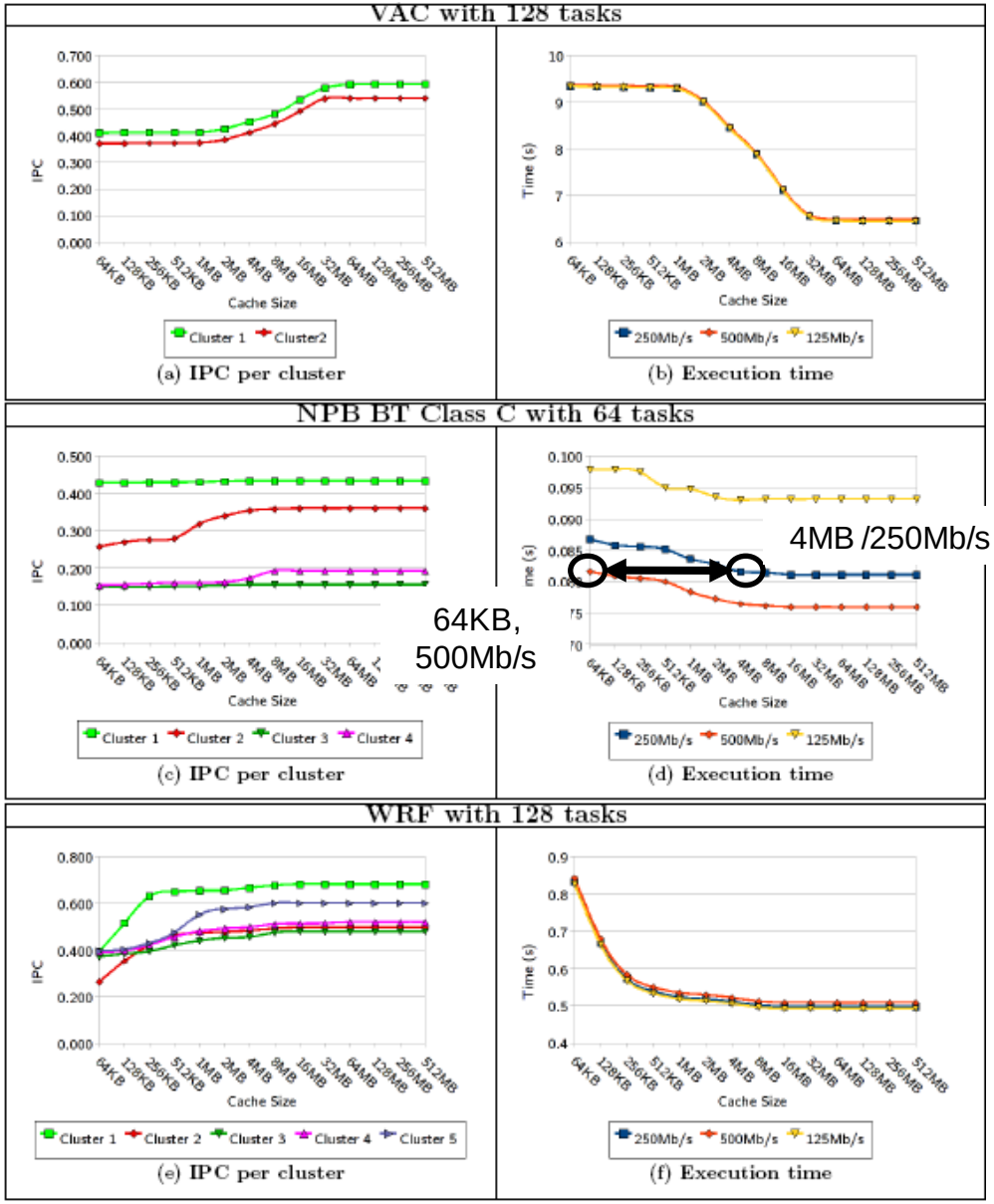






Multi-scale simulation: L2 cache size vs. Network BW

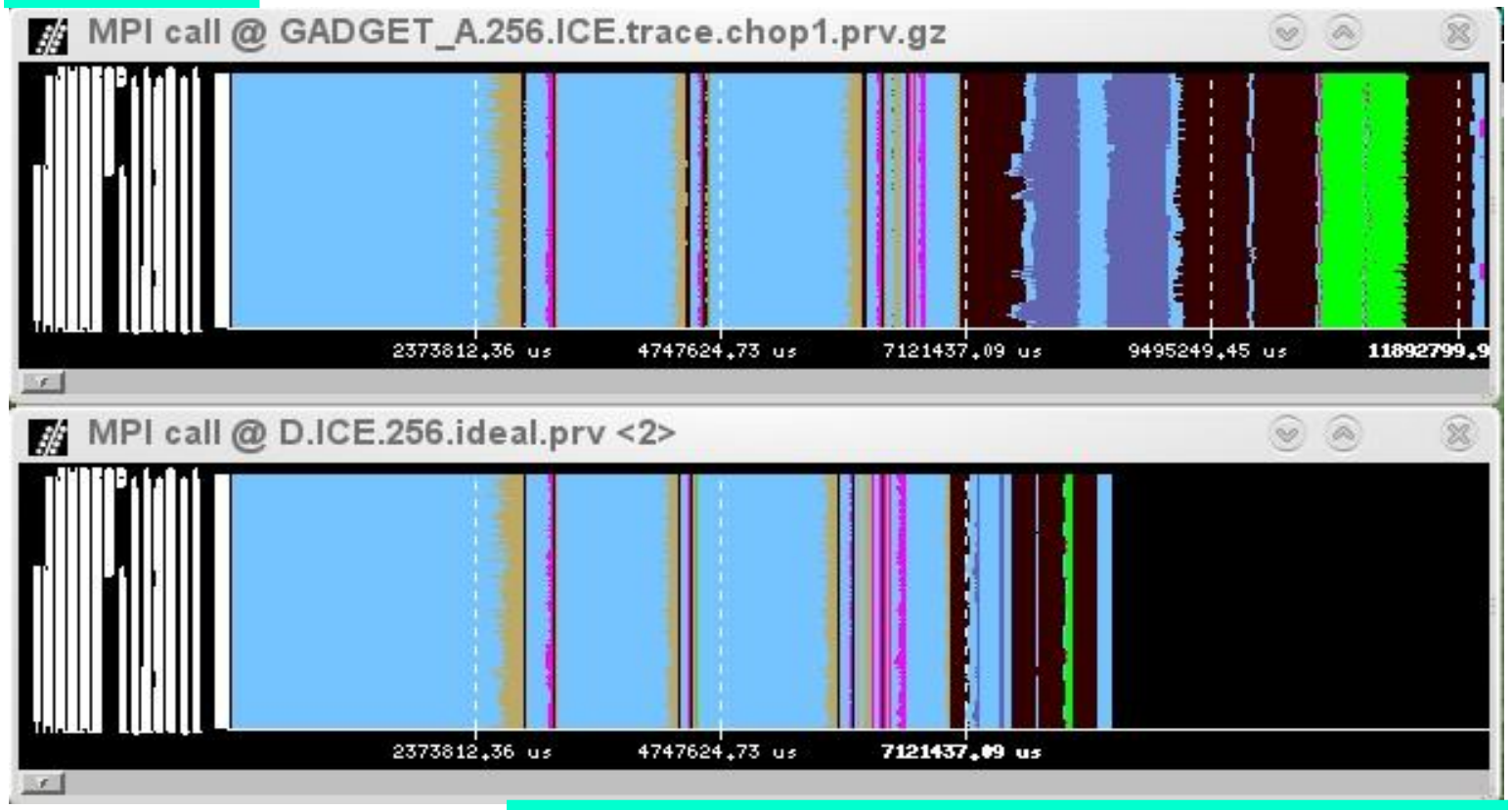
- Left: clusters IPC with different cache sizes
  - 64KB - 512MB
- Right: execution time with different network BW and cache size
  - 125MB/s – 500MB/s
- NAS BT
  - Can compensate cache reduction with more network BW
- VAC, WRF
  - Dominated by computation



GADGET: Application limits

- Load balance / dependences?

Real run

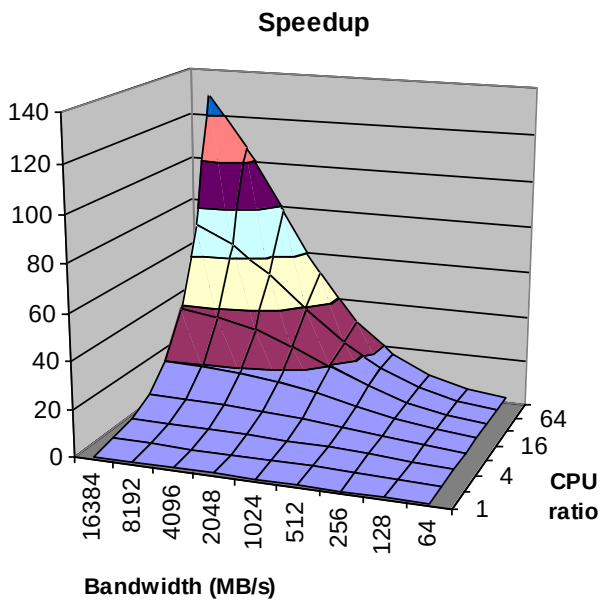


Ideal Network: infinite bandwidth / no latency

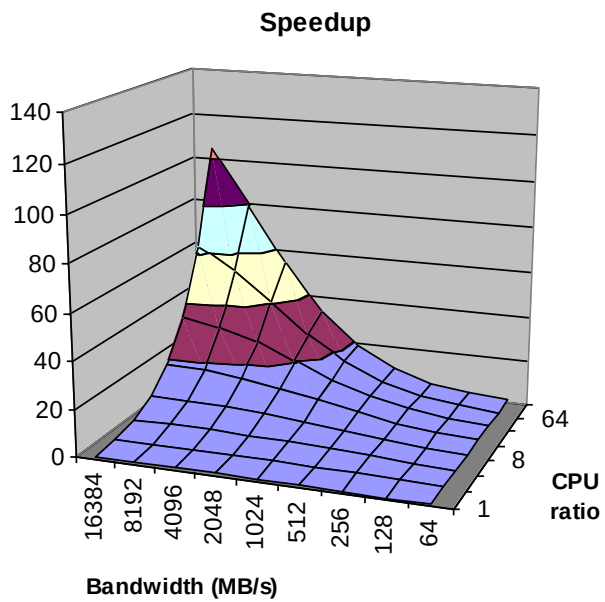
GADGET: Impact of architectural parameters

- Ideal speeding up ALL computation bursts by a the CPUratio factor
- The more processes, the less speed-up
  - Higher impact of bandwidth limitations!!!!

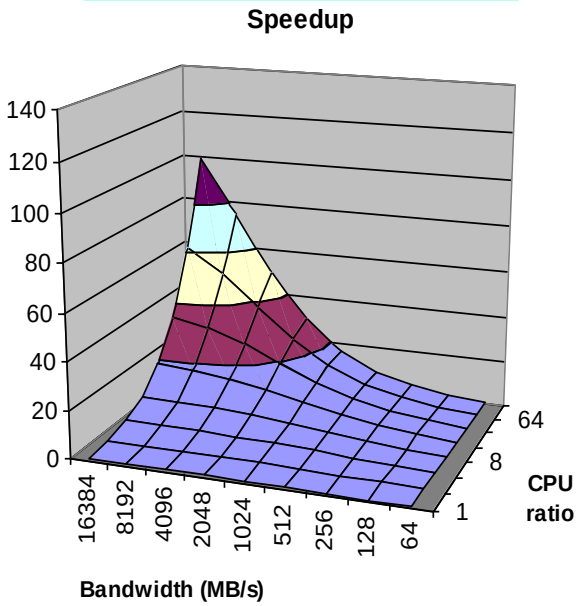
64 procs



128 procs

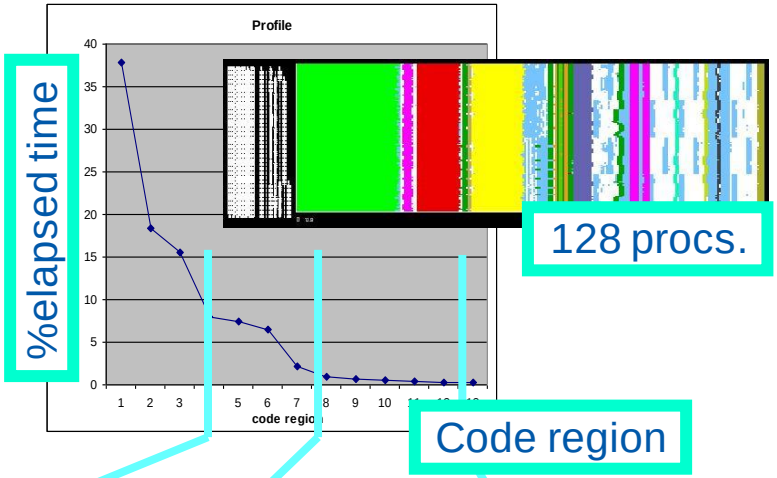


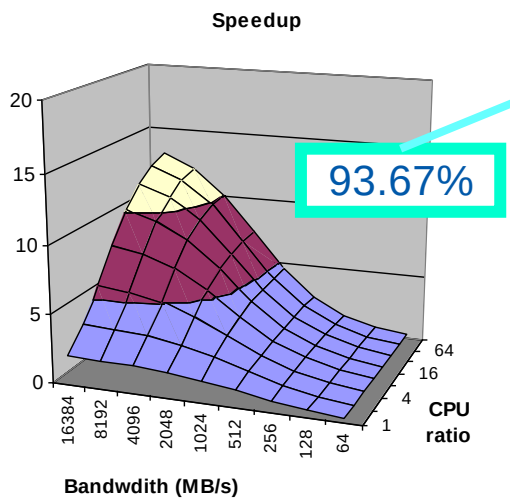
256 procs

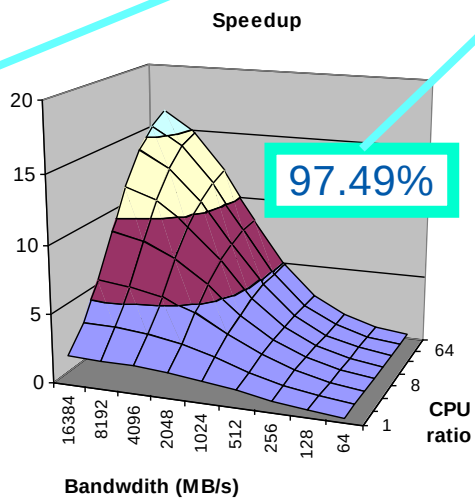


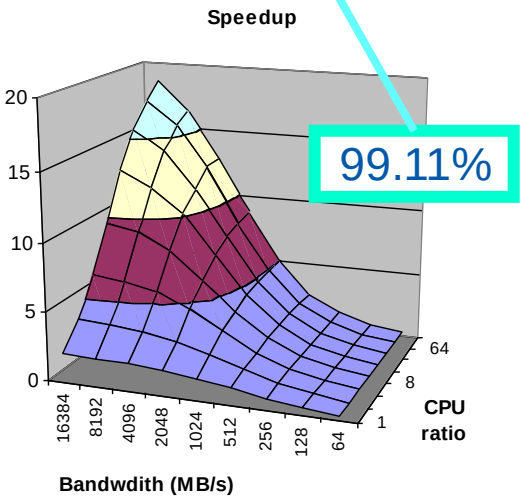
GADGET (*what-if?*): potential of hybrid parallelization

- Hybrid/accelarator parallelization
  - Speed-up SELECTED regions by the CPUratio factor
- We do need to overcome the hybrid Amdahl's law
  - Asynchrony + Load Balancing









(Previous slide: speed-ups up to 120x)

CG-POP mpi2s1D 180x120

- Comparing 10 iterations – same time scale

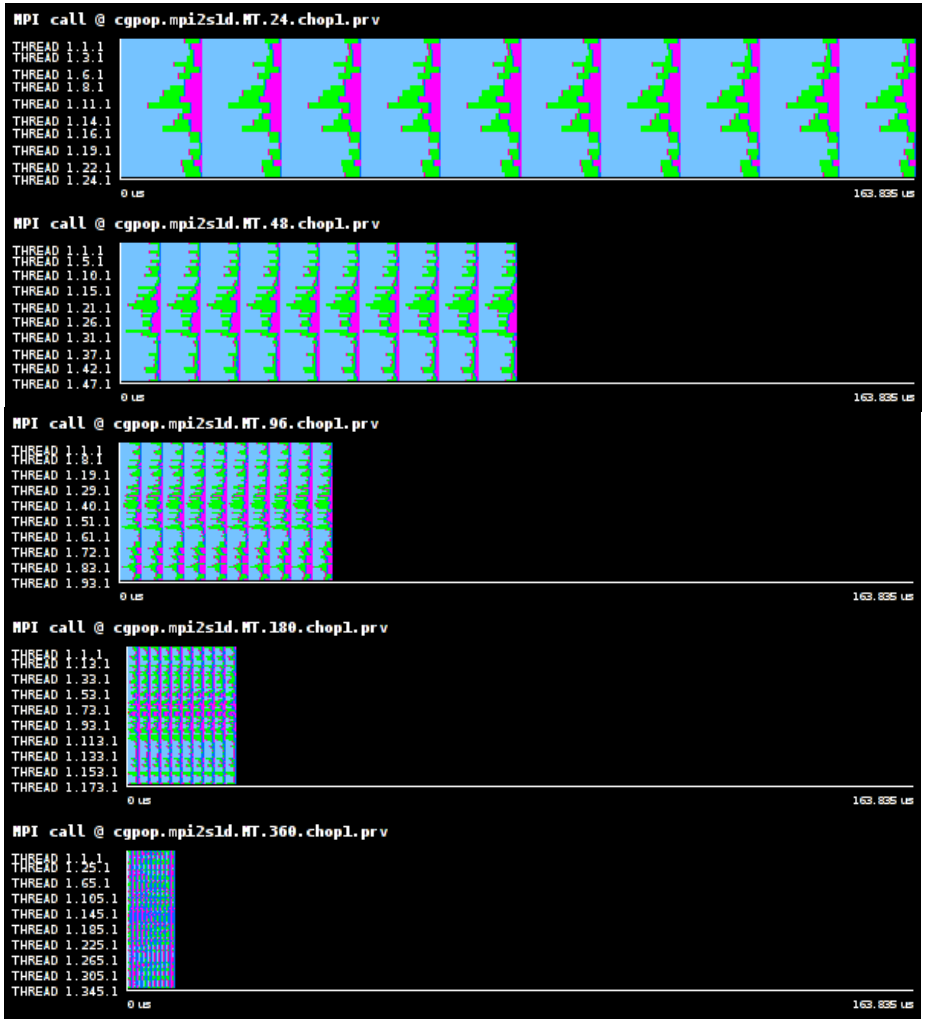
24 tasks

48 tasks

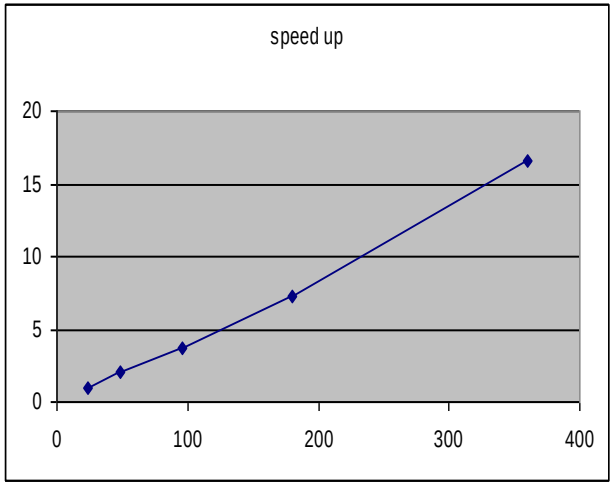
96 tasks

180 tasks

360 tasks



Good scalability!



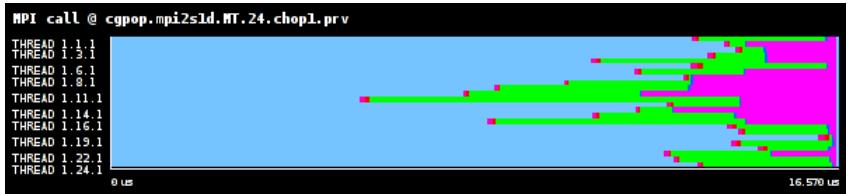
CG-POP mpi2s1D 180x120

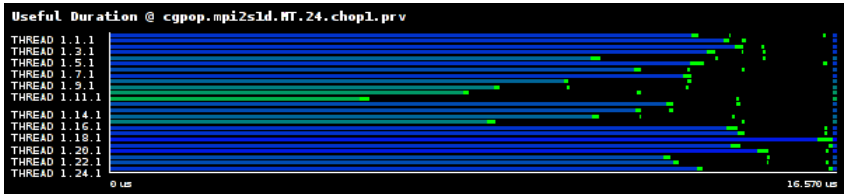
Comparing 1 iteration – different time scale

MPI call view

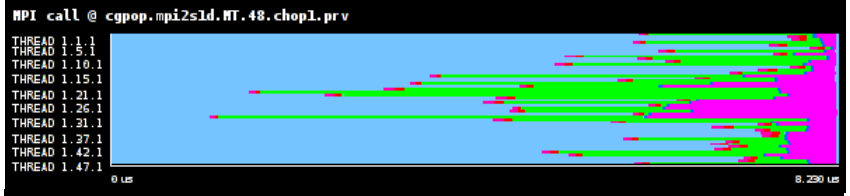
- Duration of computing phase

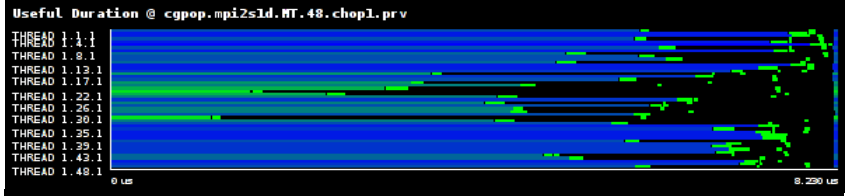
24



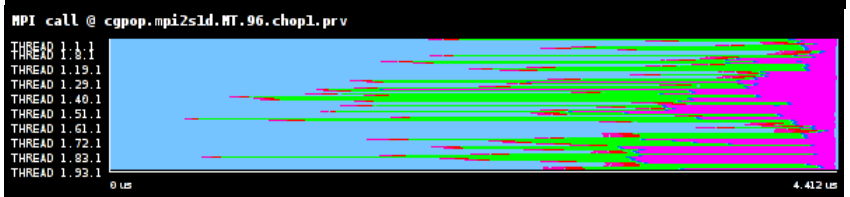


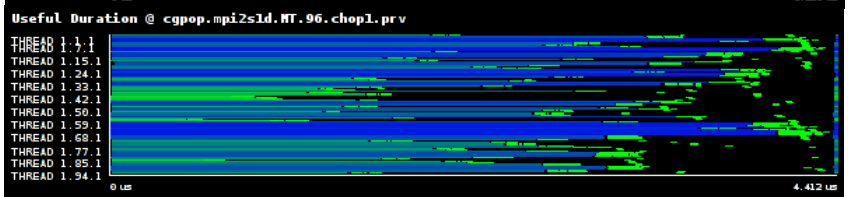
48



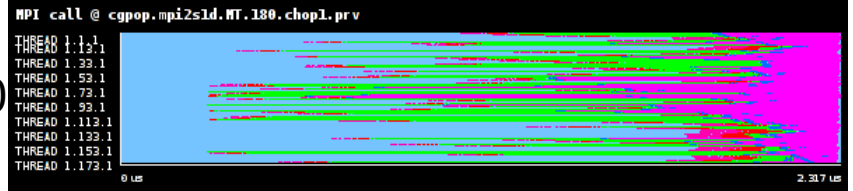


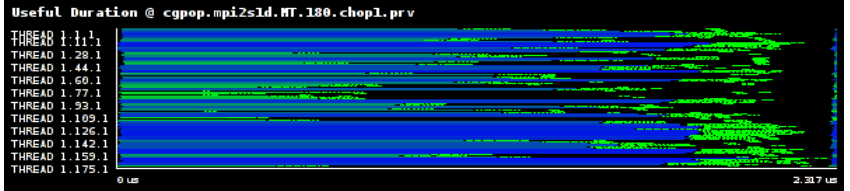
96



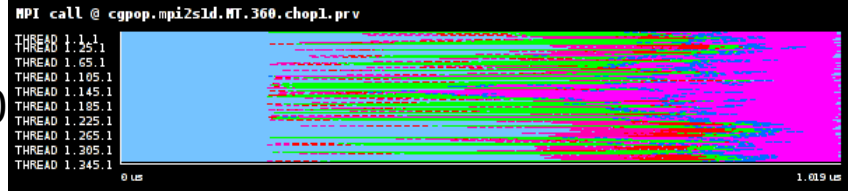


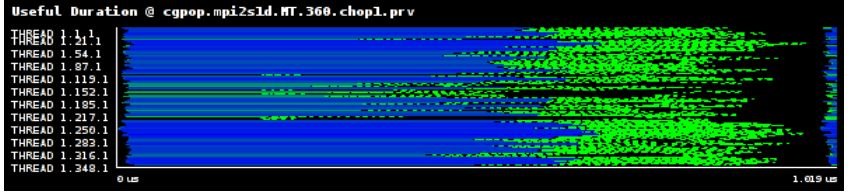
180





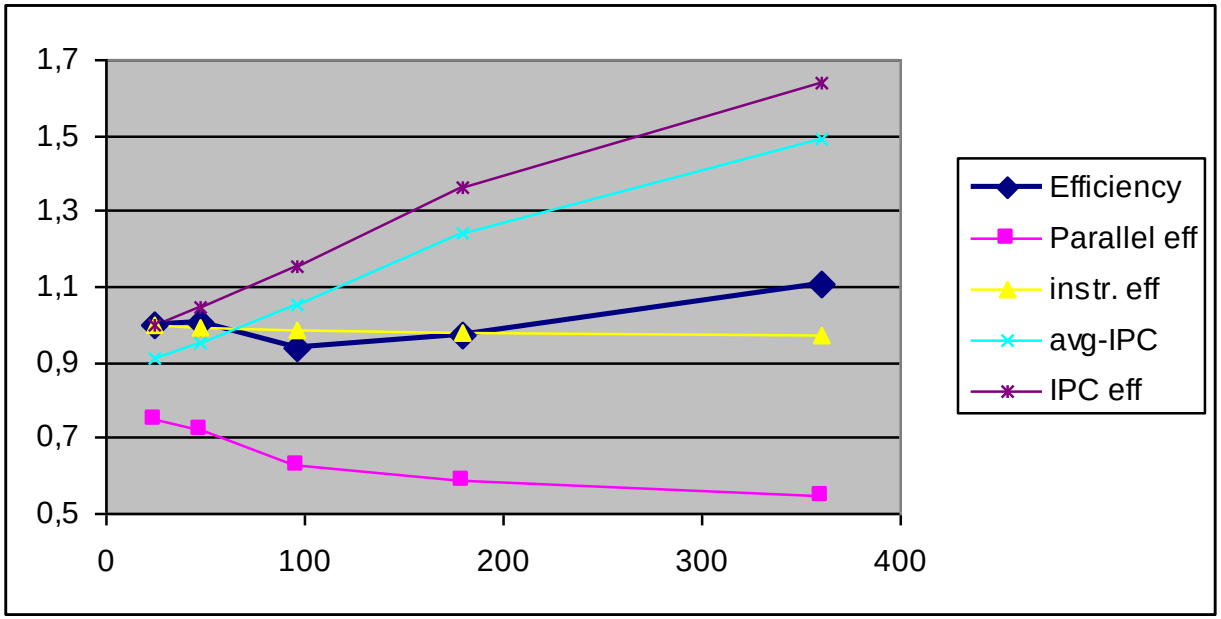
360





CG-POP mpi2s1D 180x120: Efficiency

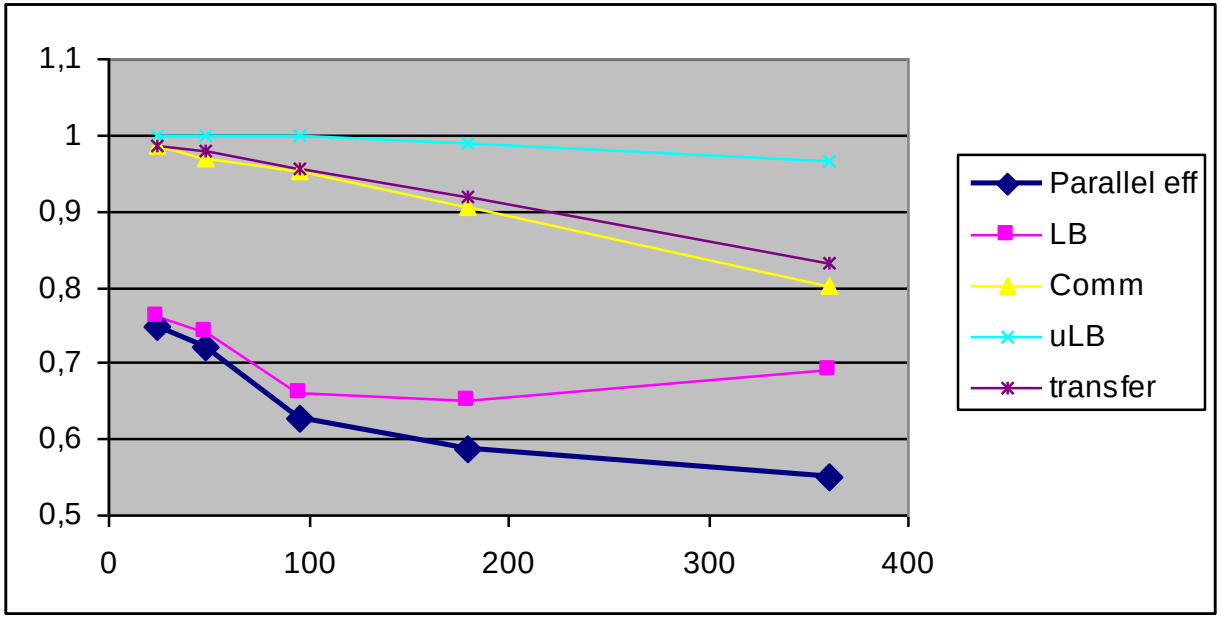
- Comparing 10 iterations program global efficiency



Cores	Time	Parallel eff.	tot-Ins	avg-IPC
24	163804479	0,75	8633761558	0,91
48	81589358	0,72	8701167887	0,95
96	43793988	0,63	8745555664	1,05
180	22508326	0,59	8834298647	1,24
360	9890676	0,55	8877820754	1,49

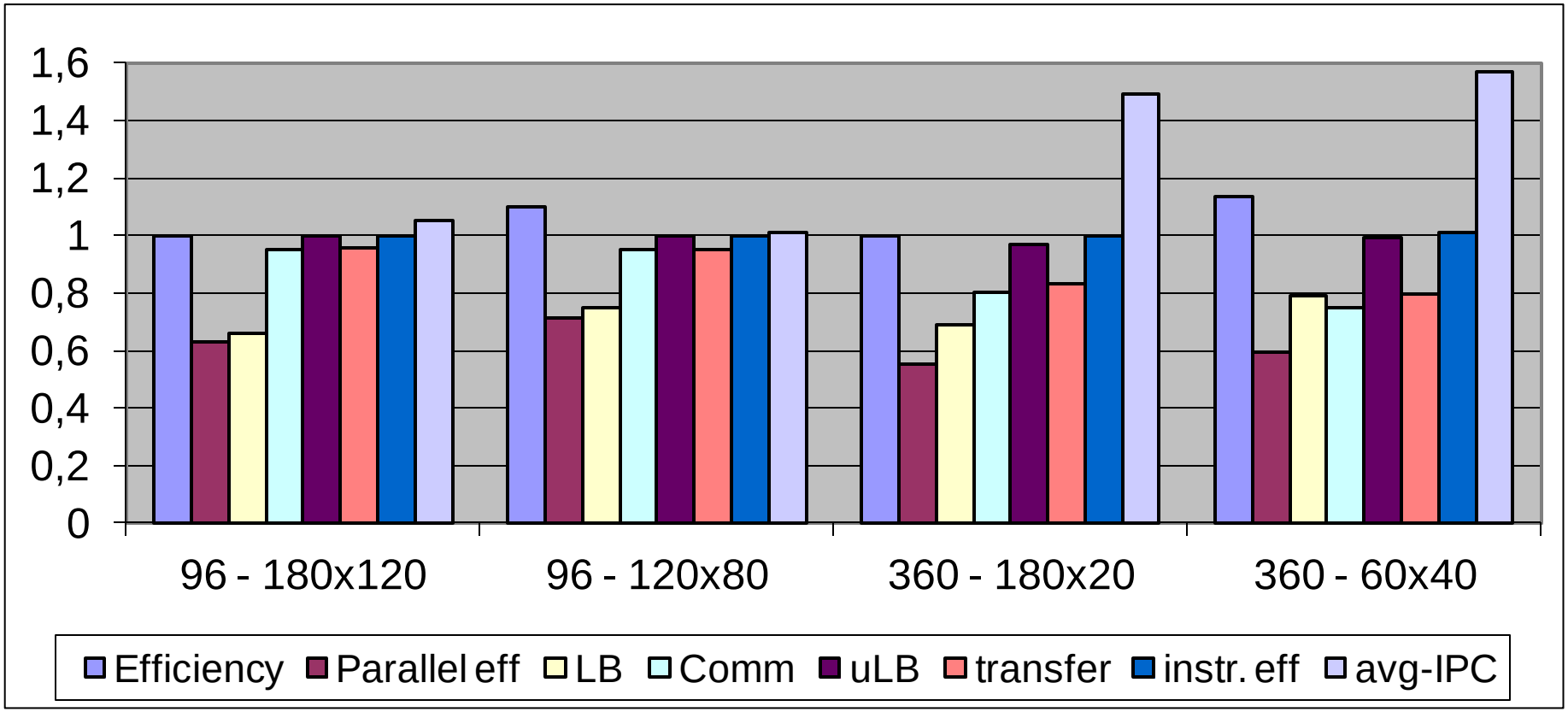
CG-POP mpi2s1D 180x120: Efficiency

- Comparing 10 iterations parallel efficiency



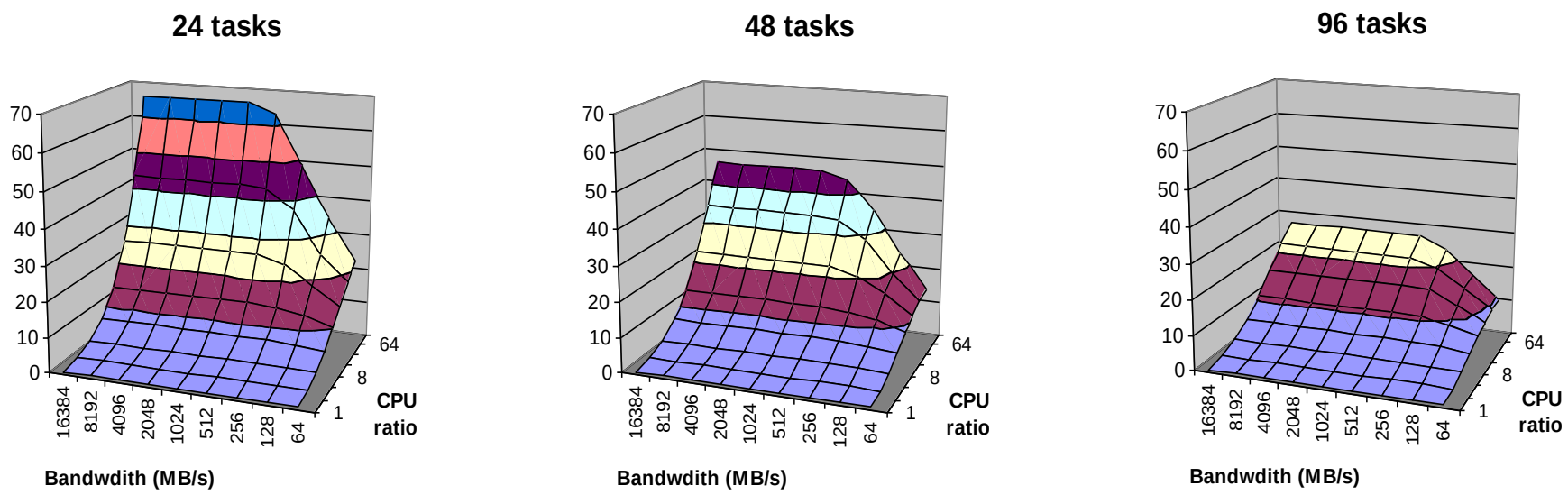
Cores	Parallel eff	LB	Comm	uLB	transfer
24	0,75	0,76	0,99	1,00	0,99
48	0,72	0,74	0,97	1,00	0,98
96	0,63	0,66	0,95	1,00	0,95
180	0,59	0,65	0,90	0,99	0,92
360	0,55	0,69	0,80	0,97	0,83

CG-POP mpi2s1D: Block size impact



CG-POP mpi2s1D 180x120: Sensitivity

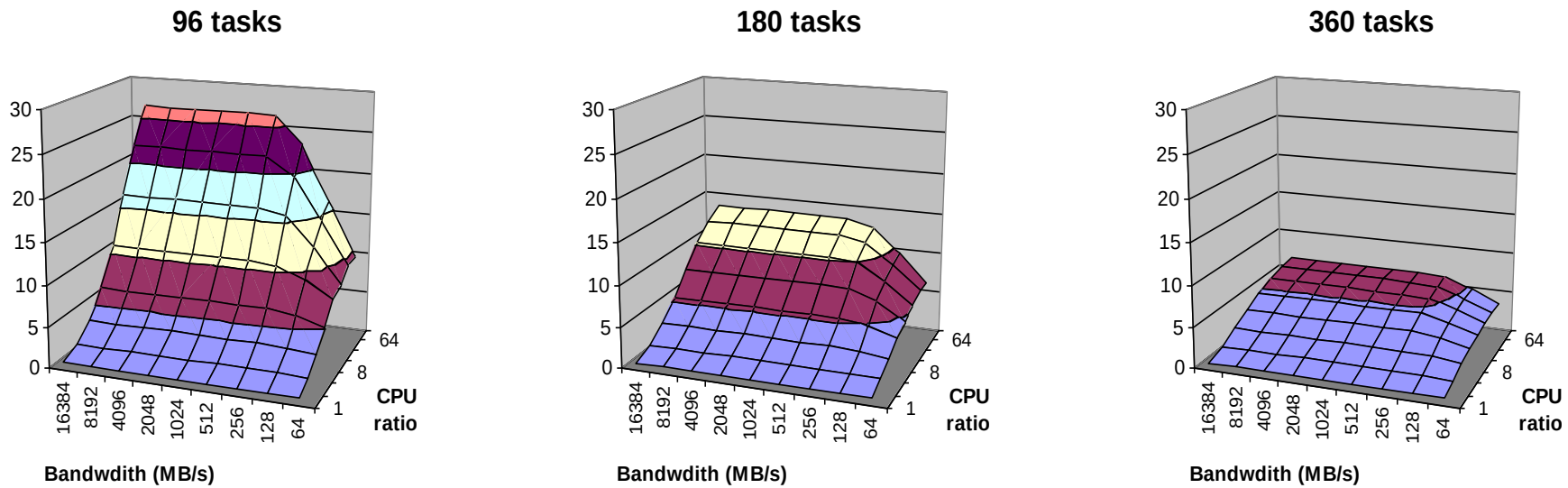
- Parametric studies
  - Sensitivity to network bandwitdh and CPU speed
  - Comparing 24, 48 and 96



  - Boundary at 512 / 1024 MB/s network BW
  - Speed-up potential reduced when scaling

CG-POP mpi2s1D 180x120: Sensitivity

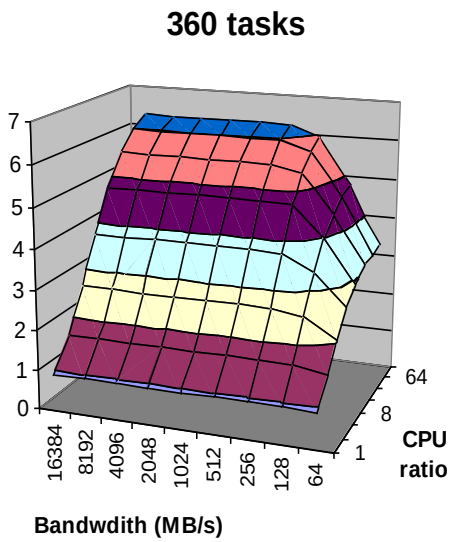
- Parametric studies
  - Sensitivity to network bandwidth and CPU speed.
  - Comparing 96, 180 and 360



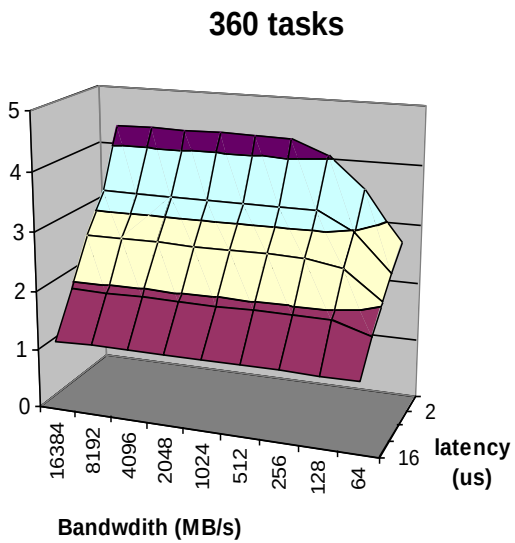
  - Boundary at 512 / 1024 MB/s network BW
  - Speed-up potential reduced when scaling
    - NOTE: previous slide max. speed-up of 70

CG-POP mpi2s1D 180x120: Sensitivity

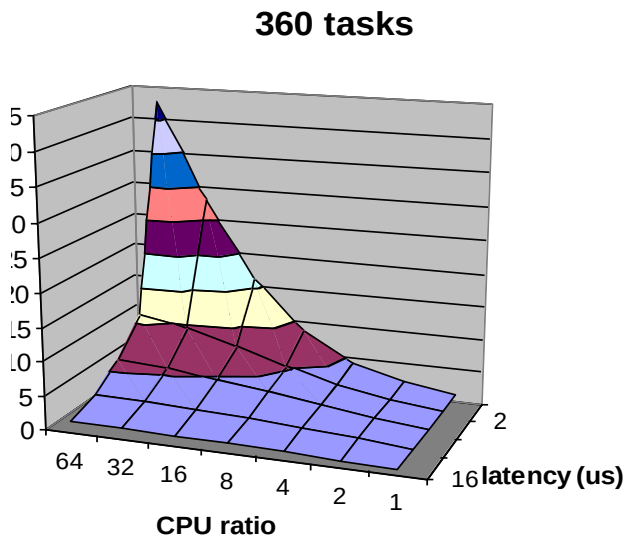
- Sensitivity to CPU, Bandwidth and latency for 360 tasks case



Latency 4 us



CPU 4 times faster

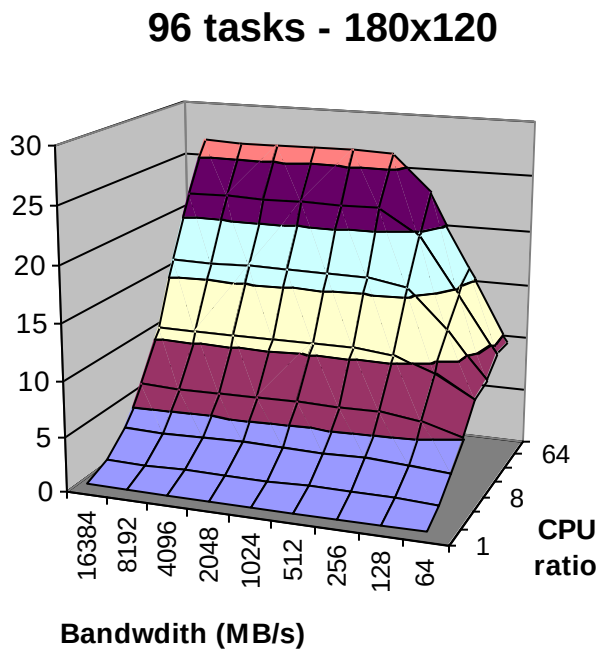


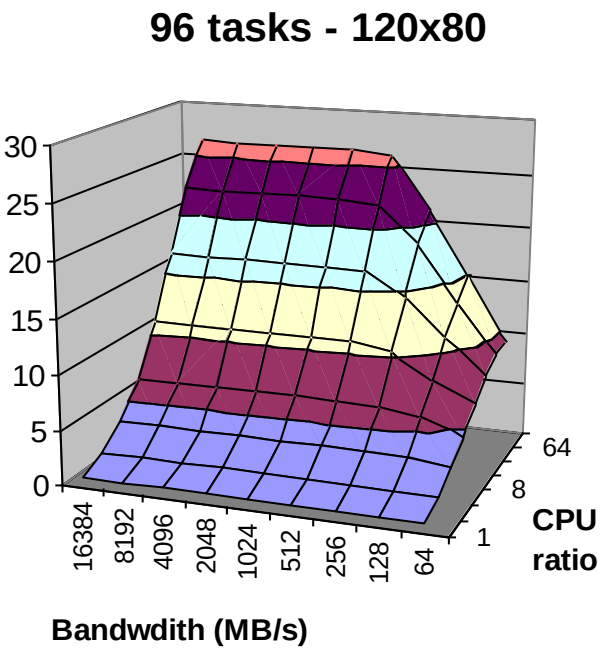
Bw 1024 MB/s

(Latency = 0 may not be realistic)

CG-POP mpi2s1D: Block size impact

- 96 tasks – Very similar

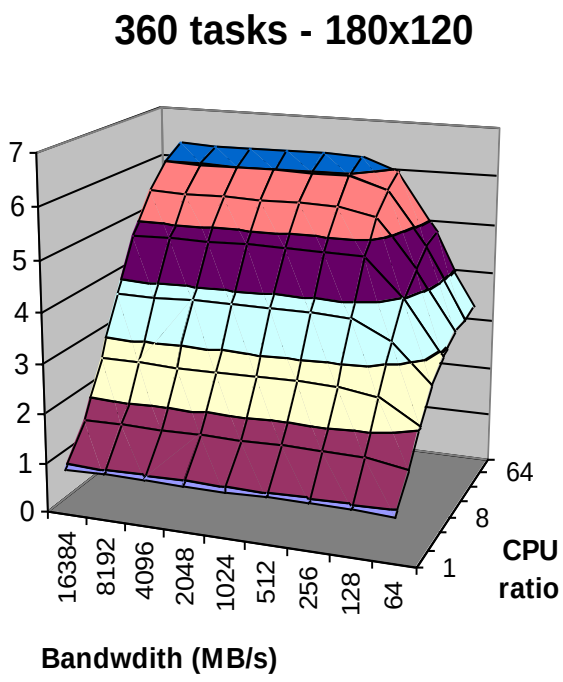


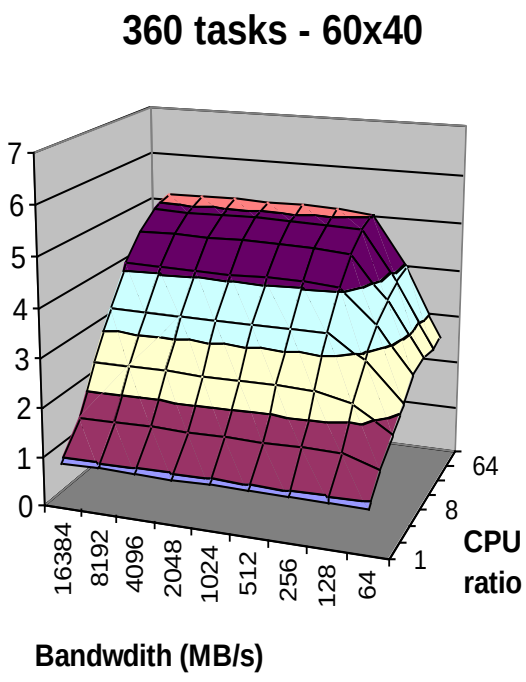


Still place for improvement?

CG-POP mpi2s1D: Block size impact

- 360 tasks – Small impact / Worst with small block size



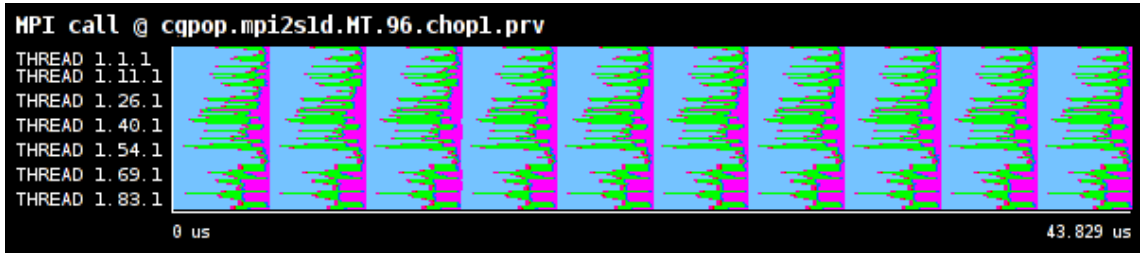


Reaching the plateau?

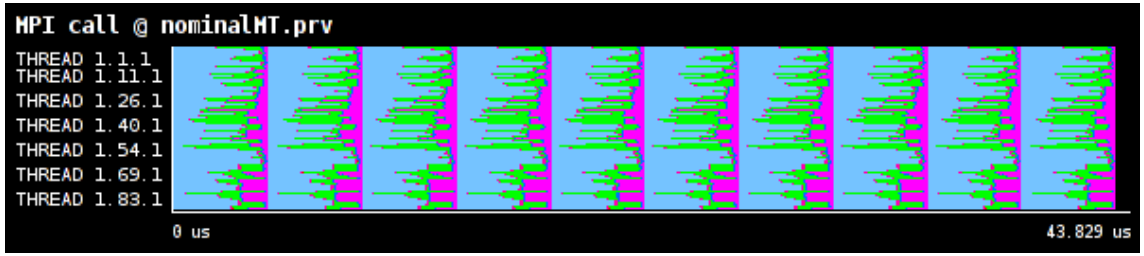
CG-POP mpi2s1D 180x120: Imbalance

- Study impact of balancing main computation
  - 96 tasks
  - Eliminate main imbalanced computation MPI calls view
  - Same time scale

Real execution



Nominal simulation

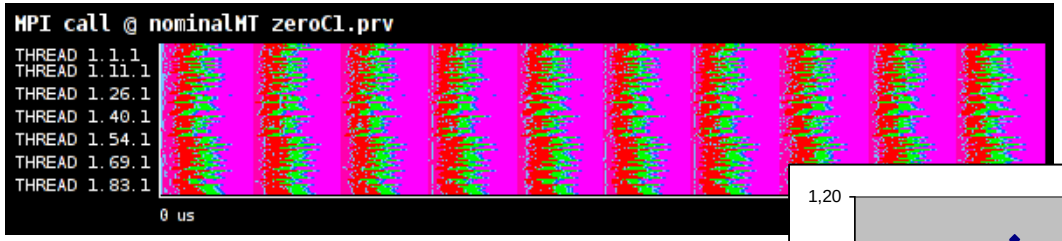


Nominal sim. Zeroing main computation

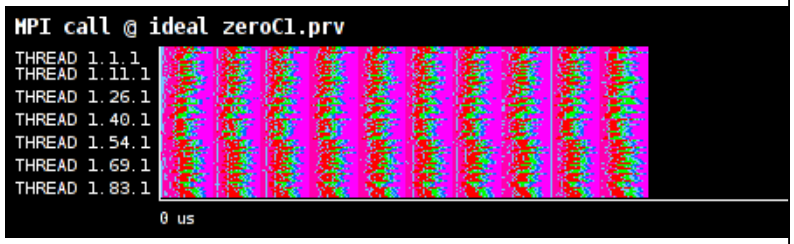
CG-POP mpi2s1D 180x120: Imbalance

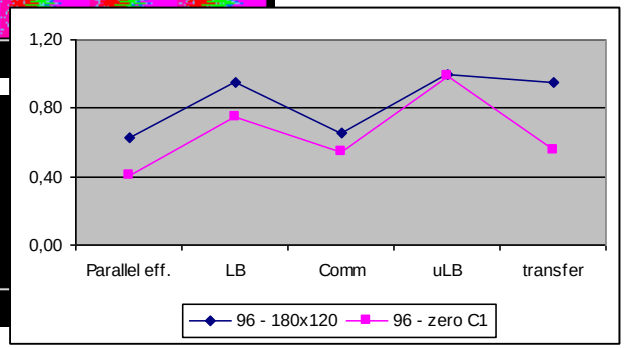
- 96 tasks
- Communication phase only
  - Small computations point-to-point / collective calls
  - MPI calls view / same time scale

Nominal sim. Zeroing main computation

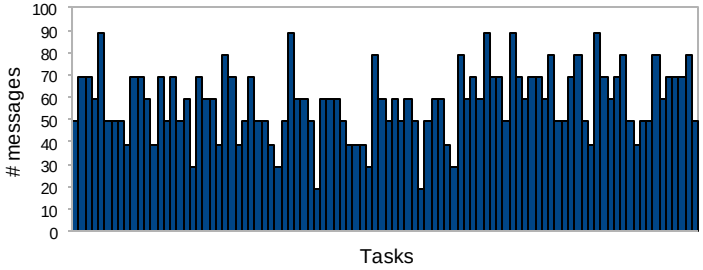


Ideal network





  - Imbalance in halo exchange
    - Computation (75%) / # messages (65%)
  - Small amount of serialization
  - Large transfer time



Conclusions

- Sometimes things are not as they look like
  - Even the scalability is good, the parallel efficiency does not scale but IPC improvements compensate
- Small scale analysis can give hints to identify problems at large scale
  - IPC variability
  - Code replication
  - *What-if* analyses